Nucleic Acids Research, 2006, Vol. 34, Database issue ***D535–D539***
doi:10.1093/nar/gkj109

BioGRID: a general repository for interaction datasets

Chris Stark, Bobby-Joe Breitkreutz[1], Teresa Reguly[1], Lorrie Boucher[1,2], Ashton Breitkreutz[1] and Mike Tyers[1,2,*]

Ontario Cancer Institute, Princess Margaret Hospital, 610 University Avenue, Toronto, Ontario, Canada M5G 2M9, [1]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada M5G 1X5 and [2]Department of Medical Genetics and Microbiology, University of Toronto, Toronto, Ontario, Canada M5S 1A8

Received September 23, 2005; Revised and Accepted October 17, 2005

ABSTRACT

Access to unified datasets of protein and genetic interactions is critical for interrogation of gene/protein function and analysis of global network properties. BioGRID is a freely accessible database of physical and genetic interactions available at http://www.thebiogrid.org. BioGRID release version 2.0 includes >116 000 interactions from *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*. Over 30 000 interactions have recently been added from 5778 sources through exhaustive curation of the *Saccharomyces cerevisiae* primary literature. An internally hyper-linked web interface allows for rapid search and retrieval of interaction data. Full or user-defined datasets are freely downloadable as tab-delimited text files and PSI-MI XML. Pre-computed graphical layouts of interactions are available in a variety of file formats. User-customized graphs with embedded protein, gene and interaction attributes can be constructed with a visualization system called Osprey that is dynamically linked to the BioGRID.

INTRODUCTION

Protein interactions assemble the molecular machines of the cell and underlie the dynamics of virtually all cellular responses (1), while genetic interactions reveal functional relationships between and within regulatory modules (2). The sum of all such interactions defines the global regulatory network of the cell (3). Proteomic and functional genomics platform technologies now generate large datasets of protein and genetic interactions, but these datasets vary widely in coverage, data quality, annotation and availability (4,5). The collation of interaction data in a consistent, well-annotated format is essential for interrogation of gene function, investigation of system level attributes and benchmarking of high throughput (HTP) interaction studies. A number of interaction databases, including BIND (6), DIP (7), HPRD (8), IntAct (9), MINT (10), and MIPS (11), provide a variety of datasets and analysis tools. We have developed a biological General Repository for Interaction Datasets (BioGRID) to house and distribute comprehensive collections of physical and genetic interactions. The precursor to BioGRID was originally conceived as a laboratory information management system (LIMS) for HTP interaction data (12). The first public release of BioGRID (version 1.0; July 2002; then termed GRID) housed HTP two-hybrid and mass spectrometric protein interaction data generated from the budding yeast *Saccharomyces cerevisiae* (13). The BioGRID has since been elaborated into a resource for HTP interaction data from other species, including the nematode worm *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster* and human. In addition, the BioGRID now contains many genetic and protein interactions curated from focused studies reported in the primary literature [Reguly,T., Breitkreutz,A., Boucher,L., Breikreutz,B.-J., Hon,G., Myers,C., Parsons,A., Friesen,H., Oughtred,R., Tong,A. *et al*. (2005) Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae* (submitted)]. The BioGRID has been queried for over 38 000 000 interactions since its inception. The recent version 2.0 release of BioGRID is a fully integrated cross-species database that supports most major model organisms, with increased data content and improved functionality.

HIGH THROUGHPUT INTERACTIONS

HTP approaches to identify novel protein and gene networks have begun to augment hypothesis-driven biochemical and genetic approaches (14). These hypothesis-generating HTP techniques include the two-hybrid (2-H) method for detecting pair-wise protein interactions (15–17), mass spectrometric (MS) analysis of purified protein complexes (12,18), and

*To whom correspondence should be addressed. Tel: +416 586 8371; Fax: +416 586 8869; Email: tyers@mshri.on.ca

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Table 1. Total number of interactions currently housed in BioGRID for indicated species

Species	Set	No. of nodes	No. of edges	No. of sources
S.cerevisiae	HTP-PI	4478	12 994	5
	LC-PI	3099	19 744	3132
	HTP-GI	1440	6119	21
	LC-GI	2656	11 234	3581
	Total	5370	50 091	5794
D.melanogaster	HTP-PI	6840	21 944	2
	LC-GI	1312	9164	1398
	Total	7216	31 108	1400
C.elegans	HTP-PI	2801	4453	1
H.sapiens	LC-PI	6374	30 761	11 921
All interactions	Total	21 761	116 413	19 116

Interactions include self edges, multiple sources and multiple experimental evidence types. HTP, high throughput; LC, literature curated; PI, protein interactions; GI, genetic interactions.

the synthetic genetic array (SGA) and molecular barcode (dSLAM) methods for systematic detection of synthetic lethal genetic interactions (19,20). BioGRID currently includes HTP protein interaction datasets from two systematic mass spectrometric studies (12,18) and three two-hybrid studies (15–17) in *S.cerevisiae,* which total 12 994 interactions between 4478 proteins (Table 1). In addition, BioGRID contains all extant HTP genetic interaction datasets from both SGA and dSLAM approaches (19–22), totaling 6119 interactions between 1440 genes. Finally, BioGRID incorporates large-scale HTP two-hybrid surveys for *C.elegans* (23) and *D.melanogaster* (24,25), among others.

LITERATURE-DERIVED INTERACTIONS

HTP datasets are laden with false positive and negative interactions (4,5). This shortfall compromises both prediction of gene/protein function and network-level analysis. The primary literature contains a vast collection of well-validated physical and genetic interactions that, while searchable on a publication by publication basis in PubMed, are not available in a relational database. A comprehensive set of literature-derived interactions would serve as a gold standard both for HTP datasets and for automated text mining approaches, augment the predictive power of HTP data and enable a re-analysis of global network properties. Spurred on by these potential applications, significant efforts to curate interaction data from the primary literature are underway by several databases (6–11), as well as by the Gene Ontology (GO) consortium (26). We have recently manually parsed the entire *S.cerevisiae* literature for protein and genetic interactions [Reguly,T., Breitkreutz,A., Boucher,L., Breikreutz,B.-J., Hon,G., Myers,C., Parsons,A., Friesen,H., Oughtred,R., Tong,A. *et al*., submitted for publication]. This comprehensive curation effort yielded 19 744 protein interactions and 11 234 genetic interactions, all of which have been placed into BioGRID. We note that the size of this literature dataset exceeds all HTP datasets combined. BioGRID also contains imports of 10 943 literature-derived genetic interactions from Flybase (27) and 30 761 literature-derived interactions from HPRD (8). The total number of literature interactions in BioGRID currently stands at over 70 000 (Table 1). In addition to the *S.cerevisiae* literature, we have curation efforts underway for the fission yeast *Schizosaccharomyces pombe*, the fruit fly *Drosophila melanogaster* and focused aspects of the human protein interaction literature, all of which will be deposited in BioGRID.

SEARCH FEATURES

The primary method of data access for BioGRID is via the web-based search interface. Combined JavaScript, PHP and Cascading Style Sheets (CSS) enable an interface that is both easy to interpret and navigate. BioGRID is supported by all main standards-compliant web browsers. Searches may be based on a wide range of supported identifiers, including gene name, ORF name, PubMed ID and free text. All genes/proteins retrieved by the query are listed in tabular format and are internally hyperlinked to allow rapid recursive searches. The BioGRID search interface retrieves the results, compiles interaction redundancies often found in large datasets and/or in combined multiple datasets, and provides an annotation-rich results page for further investigation (Figure 1). Annotation features include descriptions of gene/protein function and GO biological process, molecular function and cellular compartment terms (26).

VISUALIZATION

As network complexity increases, tabular formats for data display quickly overwhelm human comprehension. Graphical representation of interaction networks not only enables a high density of data to be visualized but immediately conveys complex inter-relationships between graph nodes, in this case either proteins or genes. A defining feature of the GRID database is an inter-dependent visualization tool called Osprey (http://biodata.mshri.on.ca/osprey) that runs as a desktop application in Windows, Linux and OSX environments (28). The Osprey platform is a facile graphical interface to query BioGRID datasets, from which the user can build custom graphical representations of any chosen set of interactions. Osprey represents individual genes/proteins by nodes and interactions by edges that connect nodes. Additional color-coded annotation is embedded in nodes and edges to represent GO categories, experimental evidence and/or data source information. A variety of graphical layouts and toggle options afford different views of the network. The Osprey file format captures all annotation associated with each node/edge in the graph, and can thus be used as a graphical file exchange format for interaction data. User-defined datasets can be up-loaded into Osprey for annotation and integration with public datasets in BioGRID. Osprey graphs can also be saved in JPEG, PNG, SVG file formats for figure construction. Pre-computed graphical representations of the first-order interaction shell for every gene/protein in the BioGRID are included on each results page and are available for direct download (Figure 1).

DATABASE STRUCTURE AND ANNOTATION

The BioGRID web interface was developed with PHP 5.0.4 and is hosted on an Apache 2.0 web server at our primary mirror (http://www.thebiogrid.org). The entire package is capable of running on any PHP 4.x compatible web server, and has been tested successfully on IIS, Apache 1.3 and Apache 2.0. BioGRID currently uses freely available

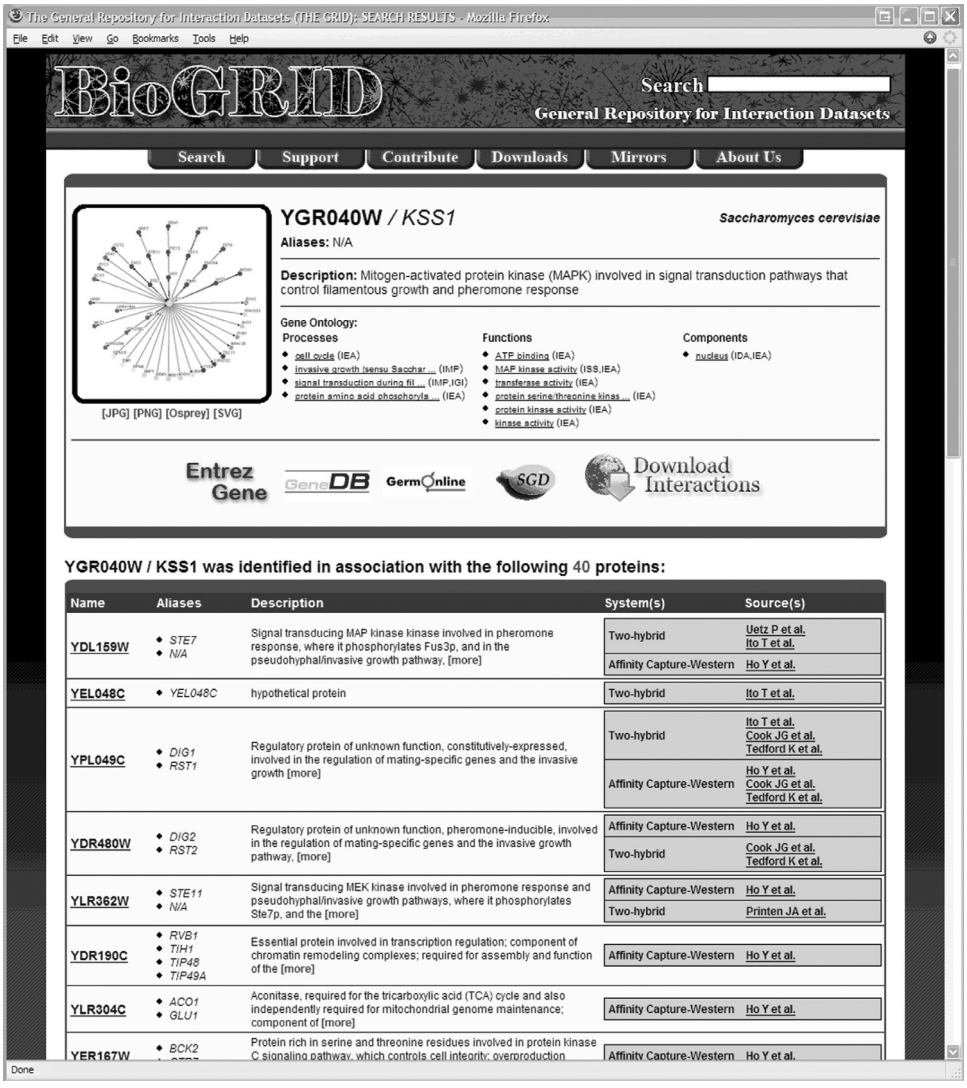

Figure 1. A sample search and result page provided by the BioGRID for the query yeast gene *KSS1*. Annotated results are collapsed to remove redundancy and hyperlinked to allow recursive searches and access to external resources. Graphical representation at upper left shows all interactions annotated with color-coded GO terms and experimental evidence. Graphs are generated by Osprey and may be downloaded in JPEG, PNG and SVG formats.

MySQL 4.1 as its primary database management system (http://www.mysql.com) for both the web-based interface and interaction curation. The BioGRID is readily established on in-house servers and is easily adapted as an internal data management system by the individual laboratory.

Consistent annotation is essential in order to collapse redundant interactions into a single search result and ensure accuracy for queries and results. All ancillary annotation is compiled from over 25 popular web-based resources, extracted and stored via an annotation compilation system (ACS) written with Java Technology and Java SDK version 1.4.2. BioGRID annotation tables are updated on a monthly basis and made freely available via the web-based interface. The BioGRID ACS currently supports 294 140 genes in 13 different organisms: *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Canis familiaris*, *Bos taurus*, *Arabidopsis thaliana*, *Xenopus laevis*, *Takifugu rubripes* and *Danio rerio*.

DOWNLOADS AND ACCESS

All the interaction data present in BioGRID is freely downloadable at http://www.thebiogrid.org. Data is available in multiple formats including tab-delimited text file and PSI-MI XML (29), as well as in Osprey and other graphical file formats. BioGRID supports the data exchange standard PSI version 2.5, as mandated by the International Molecular Exchange Consortium (IMEx) that aims to facilitate the open distribution of interaction data (see http://imex.sourceforge.net/). Interaction data is updated regularly, and all downloadable files are refreshed to reflect the most recent changes. Download files are customizable by publication, record, organism and experimental system. To maximize performance and minimize database downtime, mirror versions of BioGRID are under construction in the US and Europe. Information on curation contributions or hosting a mirror may be obtained from the BioGRID website. Source code is freely available on request. The BioGRID is actively linked

to the Saccharomyces Genome Database (30), Flybase (27) and Germ Online (31) websites.

FUTURE DEVELOPMENT

We will continue to curate interactions from major model organisms, including human, which will be posted as monthly updates of interaction data. Annotation will be routinely updated to allow unambiguous retrieval of protein/gene names. Capability to house quantitative genetic interactions and curated post-translational modifications will be implemented in the near future. We also plan to support complex and pathway descriptions, and to enable cross-species predictions though BLAST-based alignments of orthologous networks (32). A planned open source release version of the BioGRID platform, called ProtoGRID, will simplify installation of local versions of BioGRID. Similarly, the curation management system will be released to facilitate curation of interaction data by interested groups. Finally, graphical representations will be augmented through network clustering based on user-defined attributes, including co-expression and co-localization.

ACKNOWLEDGEMENTS

We thank Jim Woodgett for generous support and advice, Rachel Drysdale and Don Gilbert for assistance in parsing genetic interactions from FlyBase; Kara Dolinski, Michael Cherry and David Botstein for helpful discussions and support at SGD; and, Russ Finley, Joel Bader, Marc Vidal, Jef Boeke, Tim Hughes and Charlie Boone for pre-publication release of large-scale datasets. L.B. is supported by a National Cancer Institute of Canada Doctoral Award with funds from the Terry Fox Foundation; M.T. is supported by a Canada Research Chair in Functional Genomics and Bioinformatics. This work was funded by a grant from the Canadian Institutes of Health Research to M.T. Funding to pay the Open Access publication charges for this article was provided by the Canadian Institutes for Health Research.

Conflict of interest statement. None declared.

REFERENCES

1. Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.
2. Hartwell,L.H., Hopfield,J.J., Leibler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.
3. Bader,G.D., Heilbut,A., Andrews,B., Tyers,M., Hughes,T. and Boone,C. (2003) Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends Cell Biol*., **13**, 344–356.
4. Bader,G.D. and Hogue,C.W. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol*., **20**, 991–997.
5. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
6. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al*. (2005) The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res*., **33**, D418–D424.
7. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res*., **32**, D449–D451.
8. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M. *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*., **13**, 2363–2371.
9. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.
10. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*., **513**, 135–140.
11. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. *et al*. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*., **32**, D41–D44.
12. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
13. Breitkreutz,B.J., Stark,C. and Tyers,M. (2003) The GRID: the General Repository for Interaction Datasets. *Genome Biol*., **4**, R23.
14. Jorgensen,P., Breitkreutz,B.J., Breitkreutz,K., Stark,C., Liu,G., Cook,M., Sharom,J., Nishikawa,J.L., Ketela,T., Bellows,D. *et al*. (2003) Harvesting the genome's bounty: integrative genomics. *Cold Spring Harb Symp. Quant. Biol*., **68**, 431–443.
15. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
16. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
17. Ito,T., Tashiro,K., Muta,S., Ozawa,R., Chiba,T., Nishizawa,M., Yamamoto,K., Kuhara,S. and Sakaki,Y. (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA*, **97**, 1143–1147.
18. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
19. Tong,A.H., Evangelista,M., Parsons,A.B., Xu,H., Bader,G.D., Page,N., Robinson,M., Raghibizadeh,S., Hogue,C.W., Bussey,H. *et al*. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.
20. Pan,X., Yuan,D.S., Xiang,D., Wang,X., Sookhai-Mahadeo,S., Bader,J.S., Hieter,P., Spencer,F. and Boeke,J.D. (2004) A robust toolkit for functional profiling of the yeast genome. *Mol. Cell*, **16**, 487–496.
21. Tong,A.H., Lesage,G., Bader,G.D., Ding,H., Xu,H., Xin,X., Young,J., Berriz,G.F., Brost,R.L., Chang,M. *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.
22. Davierwala,A.P., Haynes,J., Li,Z., Brost,R.L., Robinson,M.D., Yu,L., Mnaimneh,S., Ding,H., Zhu,H., Chen,Y. *et al*. (2005) The synthetic genetic interaction spectrum of essential genes. *Nature Genet*., **37**, 1147–1152.
23. Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T. *et al*. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.
24. Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.
25. Stanyon,C.A., Liu,G., Mangiola,B.A., Patel,N., Giot,L., Kuang,B., Zhang,H., Zhong,J. and Finley,R.L.Jr. (2004) A *Drosophila* protein-interaction map centered on cell-cycle regulators. *Genome Biol*., **5**, R96.
26. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
27. Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res*., **33**, D390–D395.

28. Breitkreutz,B.J., Stark,C. and Tyers,M. (2003) Osprey: a network visualization system. *Genome Biol*., **4**, R22.
29. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al*. (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat*. *Biotechnol*., **22**, 177–183.
30. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al*. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res*., **32**, D311–D314.
31. Wiederkehr,C., Basavaraj,R., Sarrauste de Menthiere,C., Hermida,L., Koch,R., Schlecht,U., Amon,A., Brachat,S., Breitenbach,M., Briza,P. *et al*. (2004) GermOnline, a cross-species community knowledgebase on germ cell differentiation. *Nucleic Acids Res*., **32**, D560–D567.
32. Sharan,R., Suthram,S., Kelley,R.M., Kuhn,T., McCuine,S., Uetz,P., Sittler,T., Karp,R.M. and Ideker,T. (2005) Conserved patterns of protein interaction in multiple species. *Proc*. *Natl Acad*. *Sci*. *USA*, **102**, 1974–1979.